微博城市投诉文本中地理位置实体的完整性研究*

孙 赫[1,2] 李淑琴[2] 吕学强[1,2] 刘克会[3,4]

[1](北京信息科技大学网络文化与数字传播北京市重点实验室 北京 100101)

[2](北京信息科技大学计算机学院 北京 100101)

[3](北京理工大学管理与经济学院 北京 100081)

[4](北京城市系统工程研究中心 北京 100035)

摘要: 【目的】利用互动问答社区——百度知道的知识共享、更新及时的优势，弥补维护大规模地理隶属关系资源库开销大的不足，并通过百度知道自动补全缺陷地理位置实体。【方法】对缺陷地理位置实体转化为所属区域问题，并通过百度知道进行检索；根据检索结果提取特征，计算该地理位置实体属于各个区域的得分，并构建缺陷地理位置实体的所属区域特征向量；利用规则对缺陷地理位置实体进行完整化处理，实现地理位置实体完整性表示。【结果】在完整化微博城市投诉文本中的缺陷地理位置实体时，该方法的综合精确率达到 92.51%。【局限】对零地理位置实体无法完整表示。【结论】该方法对缺陷地理位置实体完整化是有效的、可行的。

关键词: 微博城市投诉文本 缺陷地理位置实体 互动问答社区 特征值计算 完整性表示

分类号: TP391.1 G35

1 引 言

近几年，随着“微博问政”的兴起，越来越多的政府部门开设官方微博与百姓互动。对于微博城市投诉信息来说，由于每天收到的投诉微博数量巨大，地理位置实体有时会缺少区域信息。一条完整的地理位置实体应包括地名区域和地名两部分，如表 1 中的微博 1。而从表 1 中微博 2–微博 6 可以看出，地理位置实体存在如下现象：地名区域缺失，如微博 2 的“中关村”；地名区域模糊，如微博 3 的“长安街”。由于地名区域缺失或模糊现象的存在，给工作人员的统计分析工作带来了极大的困难，以致于工作人员很难统计各个区域的事故发生量，从而不能及时预防事故的发生。本文将存在上述两种情况的地理位置实体统称为缺陷地理位置实体，记为 defectLoc。而且，随着时间的推移，地名及区域信息也随之变化，使得分析地名从属区域变得更加困难，如微博 3 的“崇文门新景家园”原属于崇文区，而现在属于东城区，如何及时发现地名所属区域信息的变化显得尤为重要。对地理位置实体进行完整性表示，添加缺失的区域信息，如将“中关村”规范化为“海淀区中关村”，或确定化模糊区域，如将“长安街”规范化为“东城区长安街”或“西城区长安街”，可以方便城市管理人员进行统计与分析，进一步发现地区存在的问题，实现预警功能，对以后的工作提供决策支持。

通讯作者：孙赫，ORCID：0000-0003-1133-4869，E-mail：s_hehe@126.com。

*本文系 2013 年北京市属高等学校创新团队建设与教师职业发展计划项目“大数据内容理解的理论基础及智能化处理技术”(项目编号：IDHT20130519)、北京市科学技术研究院创新工程项目“面向智慧城市的公共设施协同管理关键技术研究”(项目编号：PXM2014_17825_000002)和网络文化与数字传播北京市重点实验室开放课题“基于棋局大数据的处理及计算机博弈关键技术研究”(项目编号：ICDD201507)的研究成果之一。

表 1　城市管理投诉文本中地理位置实体示例(北京)

微博编号	微博原始内容
1	[redacted] 最近朝阳区豆各庄乡富力又一城天天晚上能闻到类似焚烧垃圾等的毒气。今天更令人气愤的是垃圾箱直接在我们楼下明目张胆地烧了。请问豆各庄乡到底要闹哪样？强烈要求@北京12345 @北京朝阳 @北京市市政市容委 找出露天违法焚烧垃圾的管理责任人和解决方案！@毛达1977 @半支烟的烟灰缸 @爱的水滴2012
2	[redacted] #城管领导听民意# @北京12345 今早的中关村，一个在我面前摆摊，一个停在我身后，从第三张照片可以看到树后面的城管车~ 城管真的管么？！
3	[redacted] 崇文门新景家园内的小餐馆，建在所谓小区人防设施上，据说各证齐全。各位看后感觉如何？小区距离长安街不足千米，早在09年就列入北京市政重点整治对象，至今未见端倪，群租成风，治安隐患深刻，物业各种不作为，环境脏乱差。提请有关部门注意！重视！@北京12345 国贸
4	[redacted] @北京12345 @朝阳区政府热线 荟康苑小区垃圾一个月无人打扫，已经垃圾成山，为了度过一个美好的春节，请帮助联系相关单位处理，谢谢。 朝阳路辅路
5	[redacted] 我发表了文章 http://t.cn/Rzw7YPm 酒仙桥梵谷水郡小区门口黑车，黑摩的乱停乱放导致交通严重拥堵，影响居民生活两年多，求相关部门进行规划管理，反映多次无效果 @朝阳区政府热线 @北京交警 @BTV财经首都经济报道 @北京朝阳 @交通新闻热线 @北京交警 @北京12345
6	[redacted] 我去走到中国音乐学院这被一辆飞驰的车掀起的尘土笼罩了，这个工地有人管吗？@北京12345

2　相关研究

目前，国内关于地名的研究多集中在识别基本地名与长度较长的复杂地名上。蔡华利等[1]抽取新闻语料中包括省、地、县、乡、村 5 级行政地理命名实体。李丽双等[2]提取人民日报语料中带有特征词(如省、市等)的地名。唐旭日等[3]采用层叠条件随机场对人民日报语料中的中文地名进行研究，并通过静态地理关系和动态地理关系，建立行政区划隶属关系。杜萍等[4]对新闻网页语料中的中文地名进行识别研究。还有其他学者[5-9]也对规范语料中简单中文地名进行识别研究。高燕等[10]利用最大熵模型对新闻报道领域最长地点实体进行识别。以上研究均是在规范语料上进行的基本地名识别，所研究的语料来自于新闻报道者，格式规范，表达统一，且其中的中文地名特征明显，易识别，如“北京市”、“山西省”等，而对于格式不规范，表达不统一的复杂地理位置实体识别效果较差。文献[11]采用分治的思想将地名识别问题转化为基本地名识别和指示词识别，在此基础上，利用词连接算法连接基本地名和指示词，最终准确识别出长度较长的复杂地理位置实体。

以上研究均集中在地名与地理位置实体的识别上，对于地理位置实体的完整性研究较少。针对缺失的区域信息的问题，相关研究多通过构建地理本体和地理知识库解决。Egenhofer[12]提出地理空间语义网络概念。杜萍[13]将地理本体与中文地名识别与抽取有机结合，主要研究消除地名歧义。地理知识库包括地名库和地名词典，如在地名库 GNIS[14-15]中共有 65 个地理类别，是一个无结构的术语表，且无任何语义关系。在地名词典 TGN[16-18]中，利用整体/部分关系，地理实体之间形成层次化的结构，同时还定义了地理实体之间的关联关系。但构建地理本体和地理知识库需要领域专家的参与，并且对已构建的地理本体和地理知识库进行一致性、完整性维护。而维护如此庞大的地理本体和地理知识库需要耗费较大的人力，并且无法及时对数据进行更新，尤其是在隶属关系发生变化时，通常需要对较多的节点进行修改，不易做到实时性。由于互动问答社区平台作为一个知识分享的资源库，每天都会有问题、答案的增添与更新，这对于及时发现地理位置隶属关系的变化提供了较大的支持。因此，本文提出基于互动问答社区——百度知道的地理位置实体完整性表示方法。

3　百度知道中地理位置实体的完整性表达

通过数据处理提取 defectLoc，再通过百度知道对 defectLoc 补全的问题进行检索，根据反馈结果提取特征，并对 defectLoc 的所属区域进行评分，构建所属区域的特征向量，利用规则对 defectLoc 进行完整性表示，使得完整化后的 defectLoc 可进行统计分析，为相关部门提供决策支持。具体流程如图 1 所示。

3.1　数据处理

利用文献[11]提出的方法进行地理位置实体识别，识别出地理位置实体后，进一步提取 defectLoc。用户在发布一条投诉微博时，除了“@北京 12345”以外，有时也会@相关区域，如表 1 中的微博 1、微博 4 和微博 5。本文根据微博@相关区域的特点，对所有投诉微博@的内容进行抽取，当@的内容存在唯一的区域信息时，微博 4 的“@朝阳区政府热线”，将该区域作为此

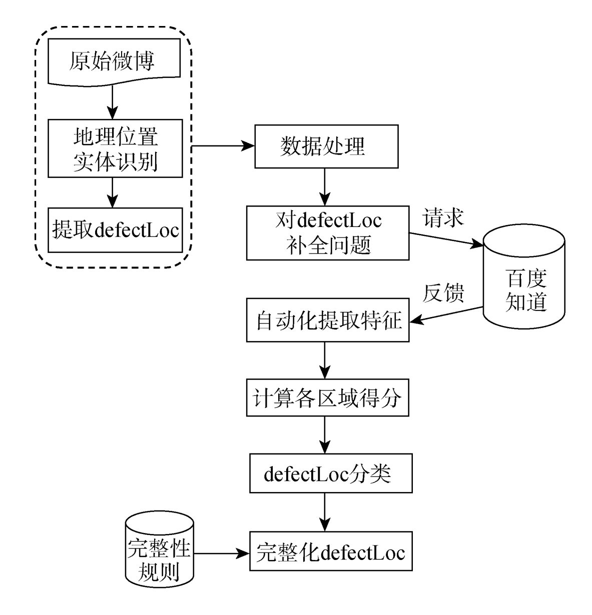


图 1　缺陷地理位置实体完整性表示框架

defectLoc 的所属区域进行完整性表示，最终过滤一部分 defectLoc。提取待处理的 defectLoc 的算法如下:

(1) 分析已识别的地理位置实体，判断其是否存在区域信息，存在则退出，如表 1 中的微博 1；不存在则转到步骤(2);

(2) 定位原微博，通过 NLPIR[19]进行原始微博的词语切分，并将所有@的内容提取出来组成@数组，判断数组中是否存在唯一区域信息，存在则补全该 defectLoc，将其过滤，如表 1 中的微博 4；不存在则转到步骤(3);

(3) 提取待处理的 defectLoc，组成 defectLoc 集合。

3.2　缺陷地理位置实体的相关问题检索

百度知道作为最流行的中文互动问答社区之一，2005 年至 2015 年 10 年间，累计解决问题超过 3.77 亿。根据文献[20]，百度知道创立后的短短两年内共产生 17 596 864 个问题，已解决 17 012 767 个问题，问题解决率高达 96.7%。同时，百度知道是一个参与率和互动性极强的知识社区，每天有超过 1 000 万用户访问，每天平均产生 71 308 个问题，223 907 个回答，平均每一个问题吸引 3.14 个用户参与互动。由于百度知道拥有大量用户群及问答数据，因此非常适合解决 defectLoc 的所属区域补全问题。

本文主要利用开放的互动问答社区——百度知道，对 3.1 节提取的 defectLoc 生成一个问题，该问题为“defectLoc 属于哪个区”，例如，“中关村属于哪个区”，通过“zhidao.baidu.com”对相关问题的检索功能，实现对 defectLoc 所属区域的搜索，将反馈结果进行结构化数据表示。如表 1 微博 2 中的“中关村”，将“中关村属于哪个区”作为检索串提交给百度知道，反馈 10 个相似问题的 QA 对集合，并对反馈的结果进行结构化表示，表 2 所示为截取的前 6 个 QA 对集合。

表 2　“中关村”结构化数据表示(部分)

排序	问题	答案	是否推荐	赞	时间
1	中关村属于什么区的?	答：海淀区得看你租什么样的房子了，还有具体的位置！一般单间的话，条件好点的 800-1200 左右，床位的话便宜些 200-400 左右！……	1	0	2009-10-02
2	中关村在北京哪个区?	答：在海淀区。	0	6	2012-02-26
3	中关村是北京哪个区的?	答：海淀区。海淀区。	0	0	2014-03-17
4	北京的中关村属于哪个区	答：中关村属于北京市海淀区管辖	0	0	2013-09-16
5	北京中关村在哪个区	答：海淀区	0	0	2015-04-11
6	请问北京市中关村是属于哪个区的?	答：中关村属于北京市海淀区管辖。	0	23	2007-06-28
…	…	…	…	…	…

3.3　缺陷地理位置实体完整性特征向量构建

根据 3.2 节的结构化数据进行反馈结果的特征提取，并对每个 QA 对的反馈特征计算其得分，通过所属区域的评分模型计算出各个区域的总得分，构建缺陷地理位置实体的得分特征向量。

(1) 百度知道反馈特征的提取

通过百度知道反馈的问答内容，本文总结三类特征，其中包括内容特征、百度知道特征和搜索反馈特征。

①内容特征

该特征描述百度知道反馈的问答内容，要确认问答内容中是否出现了区域信息。同时，如果反馈的问题与提出的问题有较高的相似度，则认为该问题、答案中出现的区域信息更重要。

1)反馈的问答对是否存在区域信息。

根据反馈的问答对构建一个 bag = {QA_1, QA_2, …, QA_{10}}, 目标区域集合为 Area={$area_1$, $area_2$, …, $area_{16}$}, 其中 QA_i 为百度知道反馈的第 i 个问答对, 每个 QA 对应一个 Area 集合。作为判断问题, 本文用十位、个位的 1 分别表示问题、答案中是否存在 $area_i$ 的区域, 如公式(1)和公式(2)所示:

$$area_i = \begin{cases} 1 & \text{QA答案中含}area_i \\ 0 & \text{QA答案中不含}area_i \end{cases} \quad (1)$$

$$area_i = \begin{cases} 10 & \text{QA问题中含}area_i \\ 0 & \text{QA问题中不含}area_i \end{cases} \quad (2)$$

针对不同区域, 每个 QA 构建一个集合包含全部区域的 16 个 area, 由于 QA 问题中出现的区域与答案中出现的区域重要性不同, 答案是对区域缺失问题的解答, 因此, 答案中的区域信息更重要。区域的得分 ScoreA 如公式(3)所示:

$$ScoreA(QA_j, area_i) = (1-\lambda)\times(area_i/10) + \lambda\times(area_i\%10) \quad (3)$$

其中, i 为第 i 个区域, j 为百度知道反馈的第 j 个问答对, λ 为答案中出现区域信息的权重。

2)问题相似度集合

这个特征用来衡量提出的问题 tq 与 QA 集合中所有问题的相似度, 记为 Simq, 则 Simq={$simq_1$, $simq_2$, …, $simq_{10}$}, 其中, $simq_1$ 至 $simq_{10}$ 为 tq 与 QA 集合中每个问题的相似度。由于余弦相似度的结果只在[0, 1]之间, 且需要计算相似度的两个问题字数较少, 通常在 10 个字左右, 因此本文采用以字为向量, 以余弦相似度作为问题相似度的计算方法。假设 A、B 是两个 n 维向量, A=[A_1, A_2, …, A_n], B=[B_1, B_2, …, B_n], 其中 A_i 与 B_i 表示同一字符分别在 A、B 中出现的频度, n 为 A、B 中所有不重复的单个字符, 则 A 和 B 的余弦相似度可以表示为:

$$simq_i = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (4)$$

②百度知道特征

百度知道特征是指百度知道本身的一些属性, 反映百度知道反馈的 QA 对的可信性, 以下特征较好地描述了 QA 对中答案的准确性。

1)是否为推荐答案

推荐答案[21]是由百度知道平台上高级知道网友推荐的质量较好的回答。因此, 推荐答案通常具有较高的可信度, 并比其他答案更加重要, 用 φ 表示推荐答案的权重。

$$Rec(i) = \begin{cases} \varphi & \text{A是推荐答案} \\ 0 & \text{A不是推荐答案} \end{cases} \quad (5)$$

2)赞次数

百度知道中, 其他用户的"赞同"可以通过竖拇指的行为对回答的准确性进行肯定, 一般赞次数越多的答案, 其质量也越高。本文对赞数的计算方式为:

$$ScoreI(QA_i, Agree) = \theta \times count(QA_i, Agree) \quad (6)$$

其中, θ 为每个赞的权值, $count(QA_i, Agree)$ 为第 i 个 QA 中的赞数。

3)回答时间

回答时间来自该 QA 对中回答问题用户发表回答的时间, 由于地理位置的区域归属问题会随时间的改变而改变, 通常越接近当前时间的回答, 其准确性越高, 因此本文对回答时间做了限制, 单位为年。

$$time_i = Now - AnsTime_i \quad (7)$$

$$ScoreT(time_i) = \begin{cases} 1 & 0 \leqslant time_i \leqslant 2 \\ 0.5 & 3 \leqslant time_i \leqslant 5 \\ 0 & time_i > 5 \end{cases} \quad (8)$$

其中, i 为第 i 个 QA, Now 为现在的时间, AnsTime 为回答问题的时间。

③搜索反馈特征

通过搜索反馈结果的顺序, 利用搜索引擎伪反馈技术[22]计算权值。由于百度知道的反馈结果中排名越靠前其与 defectLoc 的所属区域信息越相关, 本文将反馈结果的前三个查询结果看成权重相同的, 后面结果随着排名的增加权重也逐渐降低, 具体分布如公式(9)所示, 其中 i 为第 i 个 QA 对。

$$Pos(i) = \begin{cases} 1 & 1 \leqslant i \leqslant 3 \\ \frac{1}{\ln^{(i+1)}} & 3 < i \leqslant 10 \end{cases} \quad (9)$$

(2) defectLoc 所属区域的评分模型

根据是否存在区域信息、问题相似度、是否推荐、赞次数、回答时间和反馈排名的结果, 构建出每 j 条 QA 的 defectLoc 所属区域的评分模型, 其中是否存在区域信息和问题相似度作为其基数得分, 再根据不同特征的重要性, 每增加一个特征需要对已计算的得分进行修改, 如果该特征值为 0, 总得分保持不变, 反之, 特征值越大, 总得分增加的越多, 因此, 本文计算第 j 条 QA 所属于区域的得分公式如下所示:

$$\begin{aligned} RowScore(QA_j, area_i) = {} & ScoreA(QA_j, area_i) \times simq_j \times \\ & (1+Rec(j)) \times (1+ScoreI(QA_j, Agree)) \times \\ & (1+ScoreT(time_j)) \times (1+Pos(j)) \end{aligned} \quad (10)$$

综上所述, 对每条 QA 在区域的得分 RowScore 进行累加, 从而得到缺陷地理位置实体 defectLoc 属于区域的总得分 $Score(area_i \mid defectLoc)$。计算公式如下:

$$Score(area_i \mid defectLoc) = \sum_{j=1}^{10} RowScore(QA_j, area_i) \quad (11)$$

根据 defectLoc 所有区域 area 的分数值 Score，构建 defectLoc 的得分特征向量 $\{Score(area_1 \mid defectLoc), Score(area_2 \mid defectLoc), \cdots, Score(area_{16} \mid defectLoc)\}$。

3.4 缺陷地理位置实体完整性表示

根据构建出的 defectLoc 的得分特征向量，将 defectLoc 定义为以下三类地理位置实体，即明确地理位置实体、歧义地理位置实体和零地理位置实体。运用相应的规则确认一个所属区域。

定义 1 明确地理位置实体：检索结果中出现且只出现一个区域，或者 $Max(P(area_i|defectLoc)) \geq \gamma$的 defectLoc，记为 clearLoc。其中概率计算公式如下：

$$P(area_i \mid defectLoc) = \frac{Score(area_i \mid defectLoc)}{\sum_{i=1}^{16} Score(area_i \mid defectLoc)} \qquad (12)$$

定义 2 歧义地理位置实体：检索结果中出现了多个区域且 $Max(P(area_i|Location)) < \gamma$的 defectLoc，记为 ambiguityLoc。

定义 3 零地理位置实体：检索结果中未出现区域信息的 defectLoc，记为 zeroLoc。

通过对数据的观察分析发现，缺陷地理位置实体的 $Score(area_i \mid defectLoc)$ 值及检索结果中出现的区域的个数对缺陷地理位置实体的完整化起决定性作用。本文利用以下规则对不同类别的缺陷地理位置实体进行区域完整性表示。

规则 1 对于 clearLoc，存在两种情况：如果检索结果中只含有一个区域信息，则此区域信息为 defectLoc 的区域信息。如图 2 中的区域 1、4、5、7，在返回的 10 条检索结果中只有一个区域得分，也就是说在 defectLoc 构建的特征向量中，有且只有一个 $Score(area_i \mid defectLoc)$ 的分数值大于 0，且其他分数值都为 0 时，defectLoc 的区域信息为 $area_i$；如果存在 $Max(P(area_i|defectLoc)) \geq \gamma$，此 $area_i$ 为 defectLoc 的区域信息。如图 2 中的区域 6，虽然有多个区域得分，根据定义 1，即可确定所属区域。

规则 2 对于 ambiguityLoc，利用 countLoc 对 defectLoc 进行消歧。countLoc 为统计每个区域的个数，一条 QA 中出现多个相同的区域信息，按一次计算，最终得到 $Max(countLoc|area_i)$，则 defectLoc 的区域信息为 $area_i$。如果 $Max(countLoc|area_i)$存在 2 个或 2 个以上的区域，本文取第一个 $Max(countLoc|area_i)$的区域信息。如图 3 中的区域 2，“海淀”的 countLoc 为最大值 7，最终完整性规范化表示的结果为“海淀区五路居”。

规则 3 对于 zeroLoc，无法进行区域补全操作。由于此类地理位置实体不一定属于北京地区，例如图 2 中的区域 3。

		东城	西城	朝阳	丰台	石景山	海淀	门头沟	房山	大兴	昌平	顺义	通州	延庆	怀柔	密云	平谷
1	万泉庄	0	0	0	0	0	8.52	0	0	0	0	0	0	0	0	0	0
2	五路居	0.61	0	2.11	0	0	3.82	0	0	0.86	0	0	0.86	0	0	0	0
3	上虞区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	东升科技园	0	0	0	0	0	2.37	0	0	0	0	0	0	0	0	0	0
5	八王坟	0	0	11.6	0	0	0	0	0	0	0	0	0	0	0	0	0
6	前门	19	22.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	中关村	0	0	0	0	0	21.9	0	0	0	0	0	0	0	0	0	0

图 2 defectLoc 所有区域的得分特征向量

		东城	西城	朝阳	丰台	石景山	海淀	门头沟	房山	大兴	昌平	顺义	通州	延庆	怀柔	密云	平谷
1	万泉庄	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0
2	五路居	1	0	2	0	0	7	0	0	1	0	0	1	0	0	0	0
3	上虞区	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	东升科技园	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
5	八王坟	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
6	前门	6	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	中关村	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0

图 3 defectLoc 所有区域的 countLoc

对每个缺陷地理位置实体通过其所有区域得分将每个缺陷地理位置实体分类，再通过上述规则对缺陷地理位置实体进行补全，最终规范化为完整地理位置实体，如表 3 所示。

表 3　对图 2 中的缺陷地理位置实体完整性表示

缺陷地理位置实体	地理位置实体类别	完整地理位置实体
万泉庄	clearLoc	海淀区万泉庄
五路居	ambiguityLoc	海淀区五路居
上虞区	zeroLoc	上虞区
东升科技园	clearLoc	海淀区东升科技园
八王坟	clearLoc	朝阳区八王坟
前门	clearLoc	西城区前门
中关村	clearLoc	海淀区中关村

4　实验结果与分析

4.1　实验准备

实验语料来源于新浪微博，以“@北京 12345”为关键词，通过新浪微博的搜索页面“s.weibo.com”进行检索，并编写定向爬虫程序自动采集相关微博。由于投诉微博的地理位置集中在北京地区，因此，地理位置实体的所属区域包括 14 个区和 2 个县，即东城区、西城区、朝阳区、丰台区、石景山区、海淀区、门头沟区、房山区、大兴区、昌平区、顺义区、通州区、怀柔区、平谷区、密云县、延庆县。

以 1 480 条新浪城市投诉微博作为实验语料，根据文献[11]的方法共提取 1 482 个地理位置实体，并由专业人员对其进行校对。其中有 840 个地名包含明确的区域信息，可以为后续统计提供帮助，有 642 个缺陷地理位置实体，占整个语料的 43.32%。经过前期的数据处理，根据@相关区域信息的微博特点，在 642 个缺陷地理位置实体中，可以完整性表示的缺陷地理位置实体有 218 个，余下 424 个缺陷地理位置实体无法进行完整性表示。但在这 424 个缺陷地理位置实体中有 90 个重复出现过，例如“国贸”、“双井”等常见地理位置实体，去除这些重复项，总共有 334 个缺陷地理位置实体需要进行完整性表示。

通过上述数据可以看出，地理位置实体的完整性研究是有必要的，本文主要对 334 个缺陷地理位置实体进行完整性研究。经过反复实验，通常答案中出现区域信息比问题中出现区域信息对所属区域的贡献大，推荐答案对问题的解释更加权威，而百度知道特征中的赞次数对所属区域的贡献较小。针对缺陷地理位置实体，如果存在某个区域的得分超过或等于所有区域得分之和的一半时，可以确定其为明确地理位置实体，因此，本文取$\lambda=0.7$, $\varphi=5$, $\theta=0.1$, $\gamma=0.5$。

4.2　评价指标

使用精确率(Accuracy)对实验结果进行评价，即正确完整化的缺陷地理位置实体数量占全部缺陷地理位置实体的比例，其计算公式如下：

$$Accuracy=\frac{right}{total}\times 100\% \qquad (13)$$

其中，right 表示正确完整化的缺陷地理位置实体个数，total 表示待完整化的所有缺陷地理位置实体个数。

4.3　实验结果与分析

通过数据处理阶段需要对 334 个缺陷地理位置实体进行完整性表示，实验分以下三个步骤进行。

(1) 检索问题，结构化反馈结果。通过数据处理，需要对 334 个 defectLoc 进行问题检索，将问题检索的结果按照表 2 的结构进行结构化处理，最终形成 334 个反馈数据表。

(2) 特征提取，计算所有区域的得分，构建 defectLoc 的得分特征向量。采用 3.3 节的特征值计算方法及所属区域的评分模型，通过反馈数据表，计算得到每个 defectLoc 的各个区域得分，并构建出得分特征向量。

(3) 根据 defectLoc 的得分特征向量，对所有 defectLoc 进行分类，通过规则进行完整性表示。根据 3.4 节的定义，将 334 个 defectLoc 分类表示，其中有 290 个明确地理位置实体，35 个歧义地理位置实体，9 个零地理位置实体，如图 4 所示。clearLoc 约占全部 defectLoc 的 87%，说明城市投诉微博中大多数的 defectLoc 都是 clearLoc；虽然无法完整化的 zeroLoc 只约占 3%，但仍需要找到其他方法对其进行完整性表示。

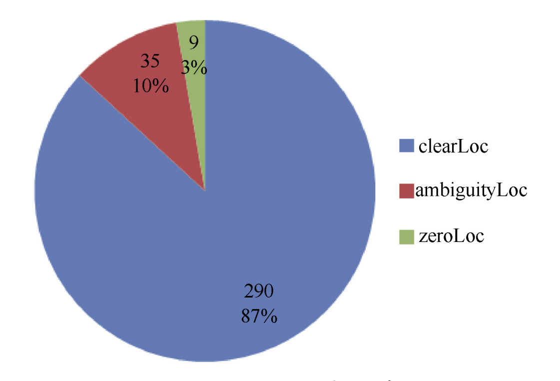


图 4　defectLoc 类别占比

利用 3.4 节中的规则对 defectLoc 进行完整化表示，

从表 4 的实验结果可以看出，本文方法完整化 clearLoc 的精确率达到 96.21%，完整化 ambiguityLoc 的精确率达到 85.71%。clearLoc 的完整化是通过 3.4 节的规则 1，由于百度知道检索得到的是唯一区域或 $Max(P(area_i|defectLoc))\geqslant\gamma$，基本不会出现歧义的区域信息，所以精确率最高。而 ambiguityLoc 的完整化精确率略低于 clearLoc，主要是存在多个歧义区域，并且得分较接近，因此在多个区域消歧过程中，有时会出现错误。本文方法可以对多数 defectLoc 实现完整性表示，覆盖率达到 97.31%。对于少数未返回检索结果的 zeroLoc，本文方法并没有效果。综上，本文方法适用于 defectLoc 的完整性表示。

表 4　缺陷地理位置实体中各类型分布表及正确率

类别	个数	错误个数	完整化精确率	综合精确率
clearLoc	290	11	96.21%	
ambiguityLoc	35	5	85.71%	92.51%
zeroLoc	9	9	0%	

如表 4 所示，其中 clearLoc 中有 11 个完整化错误，ambiguityLoc 有 5 个完整化错误，由于 zeroLoc 中的缺陷地理位置实体一部分存在所属区域，所以将 zeroLoc 均认为是完整化错误。通过对错误的分析，本文方法还存在以下问题：

(1) 较生僻的 defectLoc 通过百度知道检索出的结果不相关。如“北七佳园”，存在区域信息的问题有“上地佳园属于哪个区?”、“海淀区颐慧佳园属于什么街道?”，虽然这两个问题包含区域信息，但与“北七佳园”并无关系。

(2) 对于 zeroLoc 来说，主要有两种情况：该 zeroLoc 并不属于北京，由于地理位置实体识别阶段并不会区分是否属于北京，例如，“上虞区”并不是北京的某个地理位置；该 zeroLoc 属于北京，由于 defectLoc 较长，百度知道的相关问题较少，例如，“嘉园二里南门门口”属于北京，但反馈的检索结果并无区域信息。

5　结　语

为了能够对地理位置实体进行统计与分析，并为有关部门提供数据支撑，本文提出基于互动问答社区——百度知道的地理位置实体的完整性表达方法，通过向百度知道提问的方式对缺陷地理位置实体所属区域进行检索，根据检索结果提取特征，计算各个区域的得分，并构建所属区域的得分特征向量。在此基础上，利用完整化规则对缺陷地理位置实体进行区域补全，最终实现地理位置实体完整性表示。实验结果表明，本文方法使缺陷地理位置实体完整化具有较高的精确率，同时验证了百度知道反馈特征与完整化规则对于缺陷地理位置实体完整性表示的有效性。由于百度知道是多用户参与的互动问答社区，下一步的工作可以对零地理位置实体进行分析，利用搜索引擎、地图等多种资源相结合的方式完整化该地理位置实体。还可以将回答者作为特征进行提取，并综合多方面特征确定完整化的规则。

参考文献：

[1] 蔡华利，刘鲁，李红. 基于规则推理的突发事件发生地点识别研究[J]. 情报学报，2011，30(2): 219-224. (Cai Huali, Liu Lu, Li Hong. Rule Reasoning-based Occurring Place Recognition for Unexpected Event [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(2): 219-224.)

[2] 李丽双，黄德根，陈春荣，等. 用支持向量机进行中文地名识别的研究[J]. 小型微型计算机系统，2005，26(8): 1416-1419. (Li Lishuang, Huang Degen, Chen Chunrong, et al. Research on Method of Automatic Recognition of Chinese Place Names Based on Support Vector Machines[J]. Journal of Chinese Computer Systems, 2005, 26(8): 1416-1419.)

[3] 唐旭日，陈小荷，许超，等. 基于篇章的中文地名识别研究[J]. 中文信息学报，2010，24(2): 24-32. (Tang Xuri, Chen Xiaohe, Xu Chao, et al. Discourse-Based Chinese Location Name Recognition [J]. Journal of Chinese Information Processing, 2010, 24(2): 24-32.)

[4] 杜萍，刘勇. 基于本体的中文地名识别[J]. 西北师范大学学报：自然科学版，2012，47(6): 87-93. (Du Ping, Liu Yong. Recognition of Chinese Place Names Based on Ontology[J]. Journal of Northwest Normal University: Natural Science, 2011, 47(6): 87-93.)

[5] 李诺，张全. 利用地名用字分析的中文地名识别处理[J]. 计算机工程与应用，2009，45(28): 230-232. (Li Nuo, Zhang Quan. Chinese Place Name Identification with Chinese Characters Features [J]. Computer Engineering and Applications, 2009, 45(28): 230-232.)

[6] 李丽双，党延忠，廖文平，等. CRF 与规则相结合的中文地名识别[J]. 大连理工大学学报，2012，52(2): 285-289. (Li

Lishuang, Dang Yanzhong, Liao Wenping, et al. Recognition of Chinese Location Names Based on CRF and Rules[J]. Journal of Dalian University of Technology, 2012, 52(2): 285-289.)

[7] 李丽双，黄德根，陈春荣，等. SVM 与规则相结合的中文地名自动识别[J]. 中文信息学报，2006，20(5): 51-57. (Li Lishuang, Huang Degen, Chen Chunrong, et al. Identifying Chinese Place Names Based on Support Vector Machines and Rules [J]. Journal of Chinese Information Processing, 2006, 20(5): 51-57.)

[8] 黄德根，岳广玲，杨元生. 基于统计的中文地名识别[J]. 中文信息学报，2003，17(2): 36-41. (Huang Degen, Yue Guangling, Yang Yuansheng. Identification of Chinese Place Names Based on Statistics[J]. Journal of Chinese Information Processing, 2003, 17(2): 36-41.)

[9] 钱晶，张玥杰，张涛. 基于最大熵的汉语人名地名识别方法研究[J]. 小型微型计算机系统，2006，27(9): 1761-1765. (Qian Jing, Zhang Yuejie, Zhang Tao. Research on Chinese Person Name and Location Name Recognition Based on Maximum Entropy Model [J]. Journal of Chinese Computer Systems, 2006, 27(9): 1761-1765.)

[10] 高燕，张维维，张艳红，等. 最大熵模型在最长地点实体识别中的应用[J]. 广东石油化工学院学报，2012，22(4): 40-42. (Gao Yan, Zhang Weiwei, Zhang Yanhong, et al. Application of Maximum Entropy Model in the LLE Identification [J]. Journal of Guangdong University of Petrochemical Technology, 2012, 22(4): 40-42.)

[11] Li X W, Lv X Q, Liu K H. Automatic Recognition of Chinese Location Entity [A]. // Natural Language Processing and Chinese Computing [M]. Springer Berlin Heidelberg, 2014: 379-391.

[12] Egenhofer M J. Toward the Semantic Geospatial Web[C]. In: Proceedings of the 10th ACM International Symposium on Advances in Geographic Information System. 2002.

[13] 杜萍. 基于本体的中国行政区划地名识别与抽取研究[D]. 兰州：兰州大学，2011. (Du Ping. Study on the Ontology-Based Extraction of the Names of Chinese Administrative Division [D]. Lanzhou: Lanzhou University, 2011.)

[14] McCurley K S. Geospatial Mapping and Navigation of the Web [C]. In: Proceedings of the 10th International Conference on World Wide Web, Hong Kong, China. 2001: 221-229.

[15] Amitay E, Har'El N, Sivan R, et al. Web-a-Where: Geotagging Web Content [C]. In: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.2004: 273-280.

[16] Smith D A, Crane G. Disambiguating Geographic Names in a Historical Digital Library [A]. // Research and Advanced Technology for Digital Libraries [M]. Springer Berlin/Heidelberg, 2001: 127-136.

[17] Overell S, Magalhaes J, Rüger S M. Place Disambiguation with Co-occurrence Models [C]. In: Proceedings of the 2006 Cross Language Evaluation Forum, Alicante, Spain. 2006.

[18] Overell S E, Rüger S M. Using Co-occurrence Models for Placename Disambiguation [J]. International Journal of Geographical Information Science, 2008, 22(3): 265-287.

[19] NLPIR 汉语分词系统[EB/OL]. [2015-11-10]. http://ictclas.nlpir.org/downloads. (NLPIR Chinese Word Segmentation System [EB/OL]. [2015-11-10]. http://ictclas.nlpir.org/downloads.)

[20] 中国人知识搜索行为研究报告[R/OL]. [2015-11-10]. http://cimg3.163.com/tech/school/other/chinasearch.pdf. (Report of Knowledge Search Behavior of Chinese User[R/OL]. [2015-11-10]. http://cimg3.163.com/tech/school/other/chinasearch.pdf.)

[21] 推荐答案[EB/OL]. [2015-08-20]. http://baike.baidu.com/view/5570775.htm. (Answer [EB/OL]. [2015-08-20]. http://baike.baidu.com/view/5570775.htm.)

[22] 李学伟，吕学强，董志安，等.利用 URL-Key 进行查询分类[J]. 北京大学学报：自然科学版，2015，51(2): 220-226. (Li Xuewei, Lv Xueqiang, Dong Zhian, et al. Query Classification by Using URL-Key [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2015, 51(2): 220-226.)

作者贡献声明:

李淑琴，吕学强：提出研究命题，设计研究思路；
孙赫：设计研究方案，负责实验，分析数据，起草、撰写论文；
刘克会：论文修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储，E-mail: s_hehe@126.com。

[1] 孙赫，李淑琴，吕学强. Loc_wbtext. 被标注的地理位置实体的微博文本.

[2] 孙赫，李淑琴，吕学强. Loc.txt. 地理位置实体集合.

[3] 孙赫，李淑琴，吕学强. NoLoc.txt. 缺陷地理位置实体集合.

[4] 孙赫，李淑琴，吕学强. QA_NoLoc.rar. 缺陷地理位置实体反馈数据.

收稿日期: 2015-09-22
收修改稿日期: 2015-11-02

Retrieving Geographic Information for Micro-blog's City Complaints

Sun He[1,2]　Li Shuqin[2]　Lv Xueqiang[1,2]　Liu Kehui[3,4]
[1](Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology, Beijing 100101, China)
[2](College of Computer, Beijing Information Science and Technology University, Beijing 100101, China)
[3](School of Management and Economics Beijing Institute of Technology, Beijing 10081, China)
[4](Beijing Research Center of Urban Systems Engineering, Beijing 100035, China)

Abstract: [**Objective**] This study aims to utilize the knowledge sharing and constantly updating advantages of the Question Answering Community - Baidu Zhidao, which helps us reduce the cost of maintaining large geographical relationship resource, and find the complete location information. [**Methods**] First, we changed the incomplete location information to the approximate area names retrieved from Baidu Zhidao. Second, extracted each area's features and calculated scores of related geographic entities. Finally, we constructed the feature vectors for the areas with those geographic entities, which help us identify the geographic locations of these posts. [**Results**] The proposed method could retrieve accurate geographic information from 92.51% of City Complaints from the Micro-blog platform. [**Limitations**] The proposed method could not analyze posts without any geographic location information. [**Conclusions**] Our study found an effective and feasible way to locate the missing geographic information.

Keywords: City complaints of Micro-blog　Defect location entity　Question Answering Community(QAC)　Eigenvalue calculation　Integrity